Comparing and Understanding Data Distributions in EOS Applications

David Kao (NASA Ames Research Center)
Jennifer L. Dungan (NASA Ames Research Center)
Alex Pang (University of California, Santa Cruz)
Han-Wei Shen (Ohio State University)

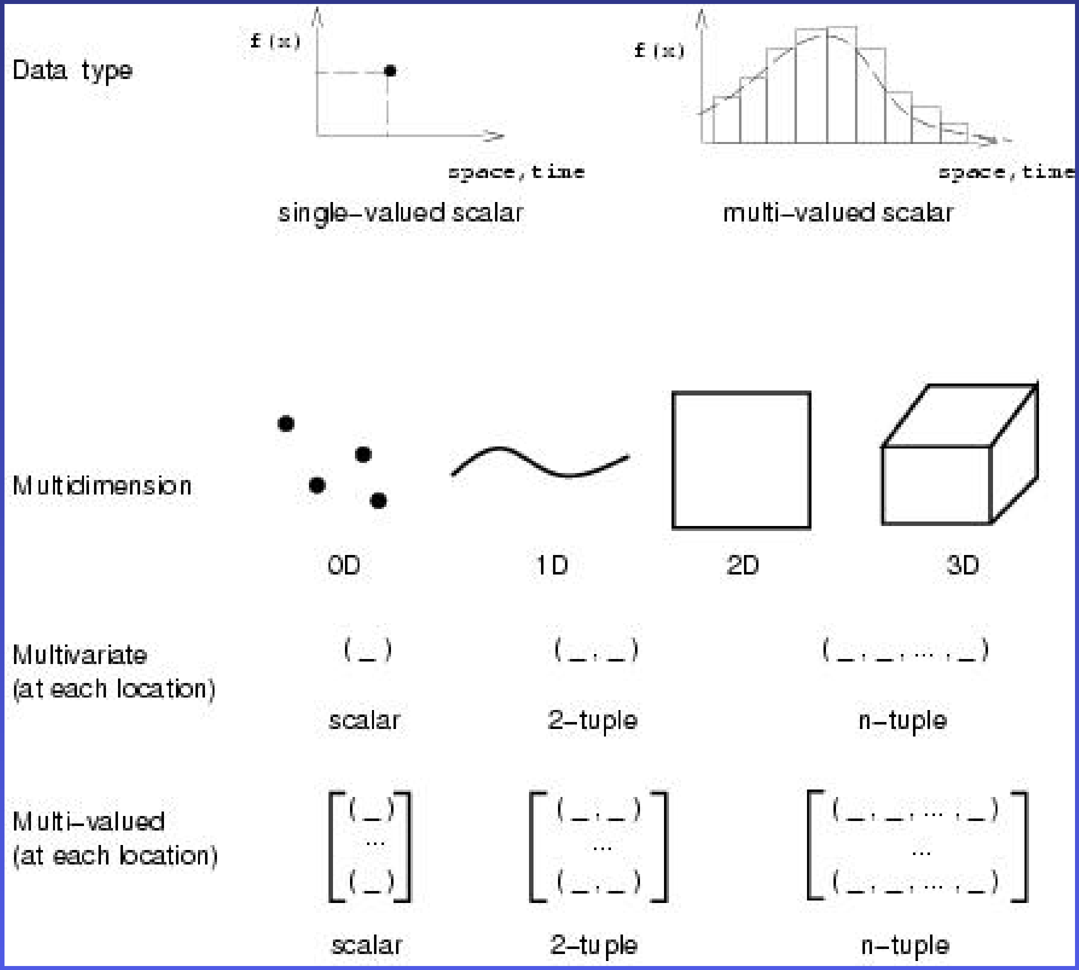
NASA Intelligent Systems Workshop
February 2004

Multi-valued data sets arise in many studies of bio and geo-physical phenomena

1. A collection of values measured at a single location, such as those from a probabilistic model.
   - Several, alternative scenarios from different models or from different parameterizations of the same model
   - Multiple, conditionally simulated realizations from a spatial process

2. A collection of values measured within an area.
   - Multiple returns of lidar pulses from each grid cell in a grid

Multi-Valued Data



We recognize this new data type, which we call multi-valued data, as a collection of values about one variable at each location.

Motivation and Goals

- Just as confidence intervals are reported for non-spatial variables, there is a need to understand the distributions of multiple values to fully represent model or measurement results.
- The goal of this project is to develop techniques to aid in the understanding and comparison of distributions from 2D multi-valued data sets.

Research Questions

- How applicable are current multivariate analyses and visualization techniques to multi-valued data sets?
- What kind of visualization techniques need to be developed to present important aspects and features of distributions of multi-valued data?
- How should the analyses and visualization tools above be further expanded to allow comparisons of distributions?
- What kind of time-varying features can we extract and exploit in helping the user understand data better?

NASA Relevance

For NASA's Earth Observing System (EOS) products and models results, common questions are:

- **Where are the values uncertain?**
- **Where are the distributions multi-modal?**
- **Where do the distributions match a target distribution?**
- **What is the area of features that answer the above questions?**

Previous Accomplishments

- For visualizing/analyzing static distribution data:
  - Used a parsimonious set of statistics for summarizing central tendency, spread and shape for pixel-wise distributions
  - Used value thresholds to define features in 2D and then generated distributions of the areas of these features. Feature-wise distributions are then summarized.
  - Summaries were rendered using standard image visualization techniques such as contour planes, surface graphs and line bars.
  - Used these techniques to capture multiple summaries in a single figure.

Previous Accomplishments

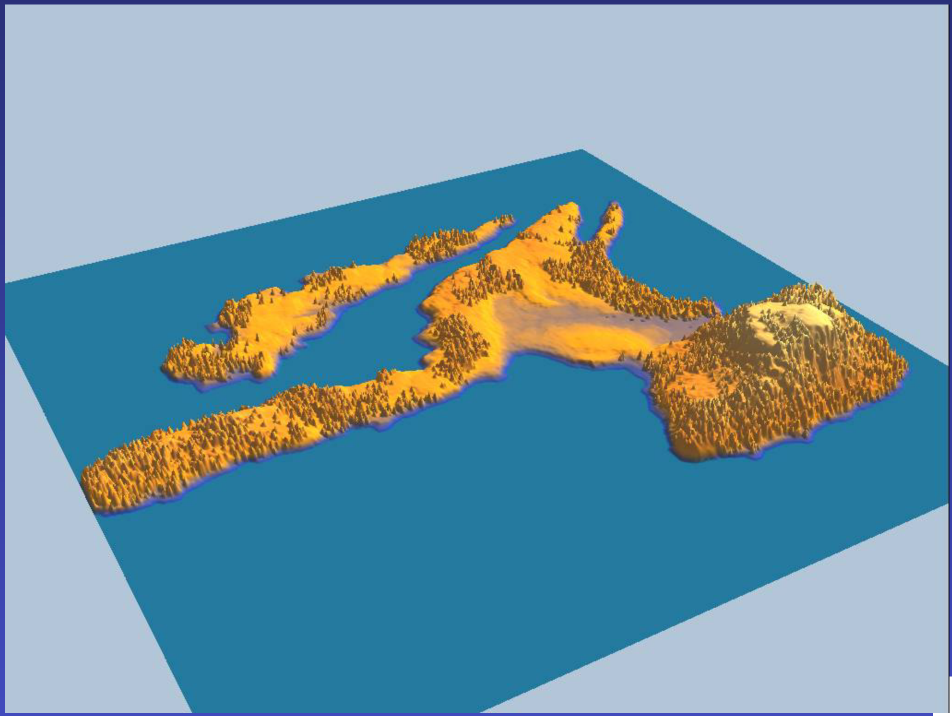
- For time-varying distributions:
  - Explored spatial and temporal clustering to describe the map's temporal variations and spatial correlations
  - Developed efficient data structures for fast temporal-spatial region indexing
  - Developed techniques to summarized the temporal behavior of the data in time intervals of arbitrary lengths

Recent Accomplishments

- Developed an operator-based approach to visualize multi-valued data sets.
- Developed distribution-matching algorithms and mode-query capabilities.
- Developed a tool, called PDFVis, that facilitates uncertainty visualization.
- Began to apply methods to spatial distributions, using lidar data as a case study. Initiated a new collaboration with a domain expert in lidar (M. Kramer).

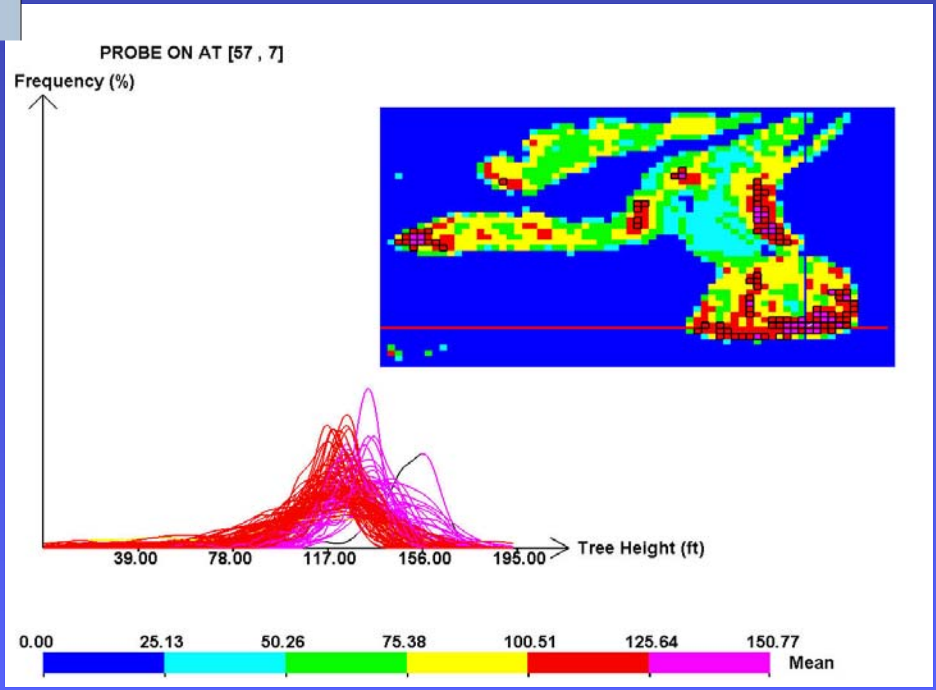
Recent Accomplishments (Cont'd)

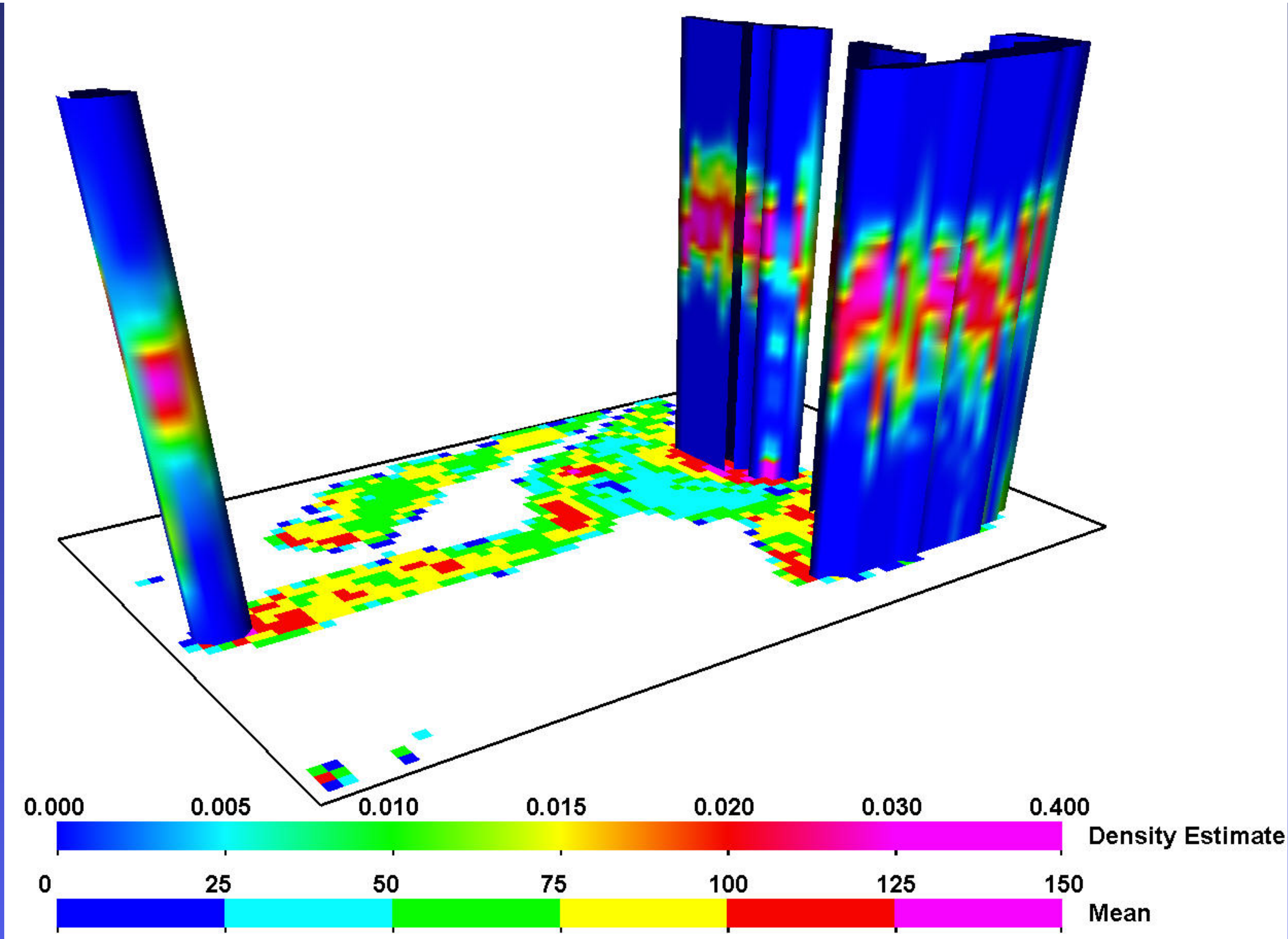
- Developed a cluster-based approach for PDF data reduction and visualization:
  - Employed hierarchical clustering of pixels
  - Investigated appropriate distance metrics for distribution data
  - Developed texture-based visualization of PDF characteristics
  - Investigated realization clustering as an alternative to pixel clustering
  - Extended the technique to spatio-temporal data



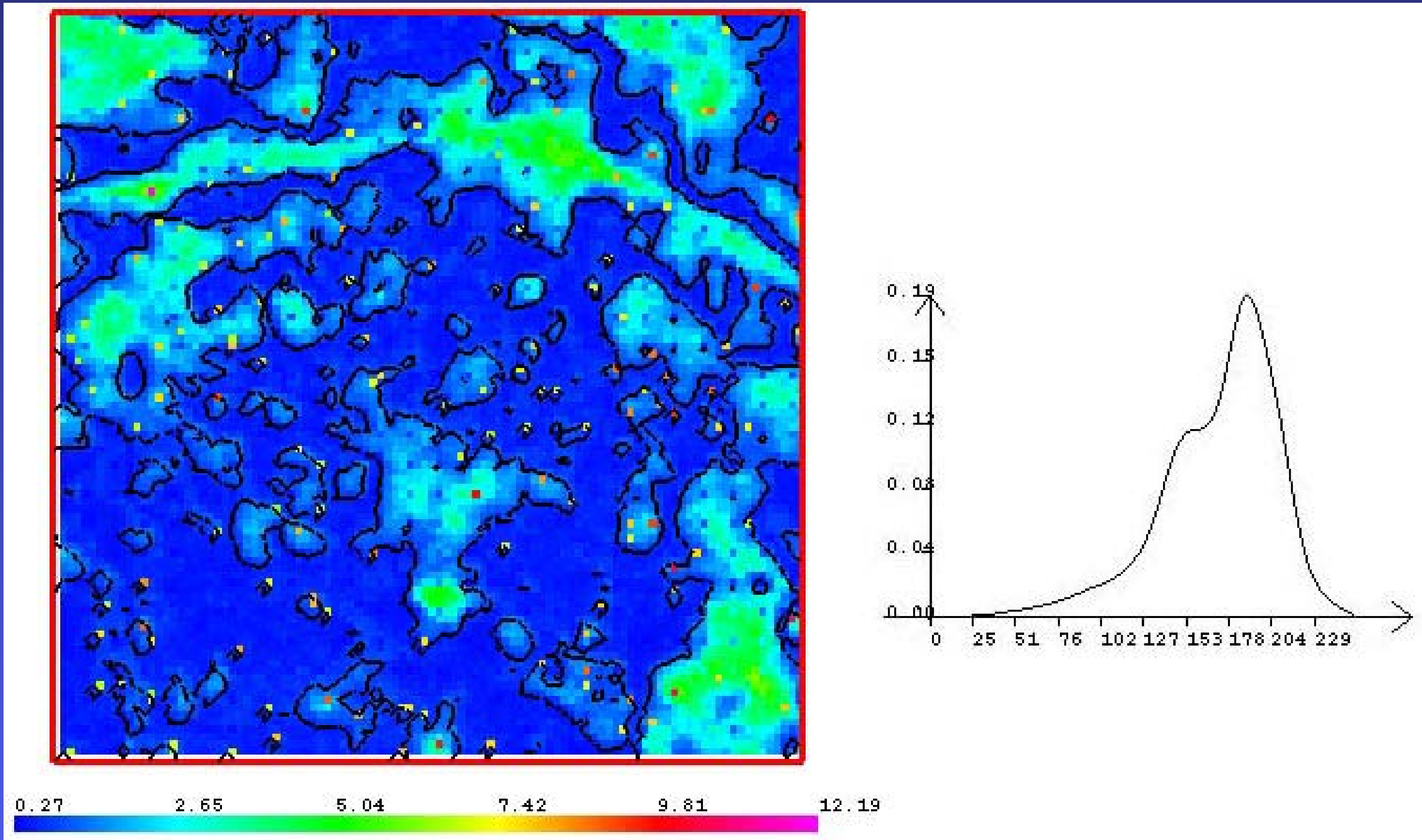
A graphical model of the High Island forest from the lidar data. (Courtesy of M. Kramer).

The graph on the left shows the distributions of those PDFs that matched a query where the distribution is unimodal and the mode occurs between 117 and 194 feet. The image on the right shows the mean tree height across the island.

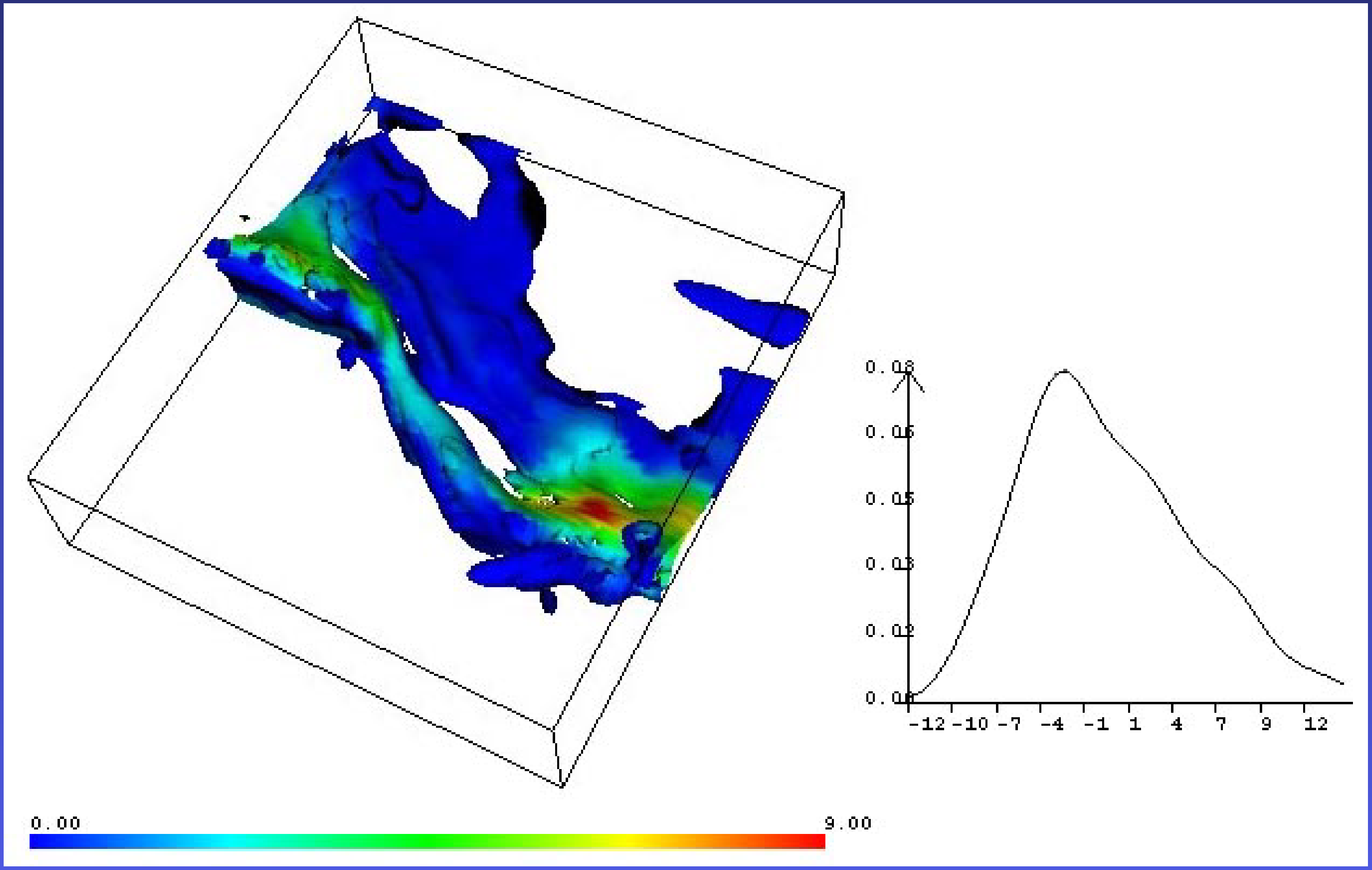




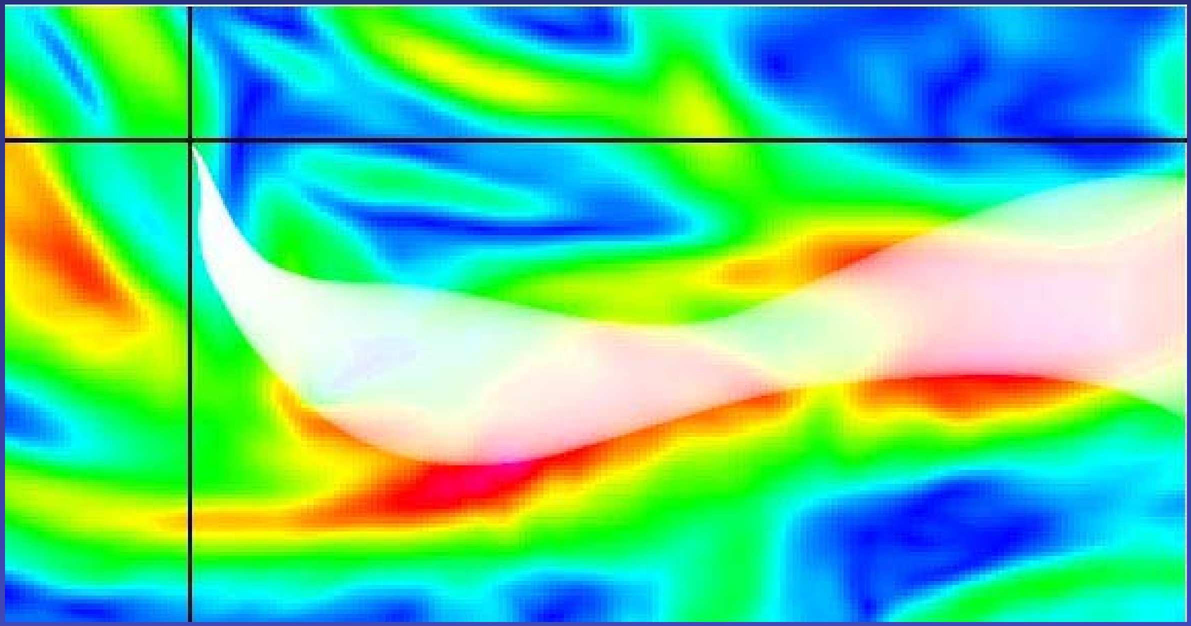
Characteristic distribution surfaces for those distributions that are found to be recovering from a recent disturbance event. The color of the surface graphs are the density estimates. The image plane is colored by mean tree height.



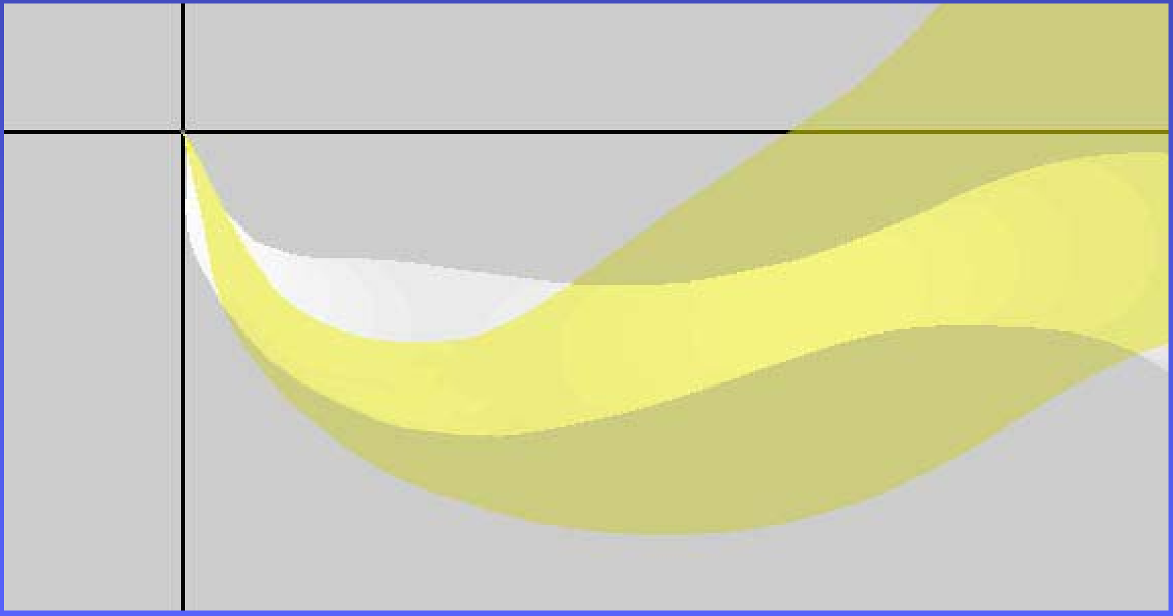
Contour lines of a target distribution.



Isosurfaces of a target distribution.

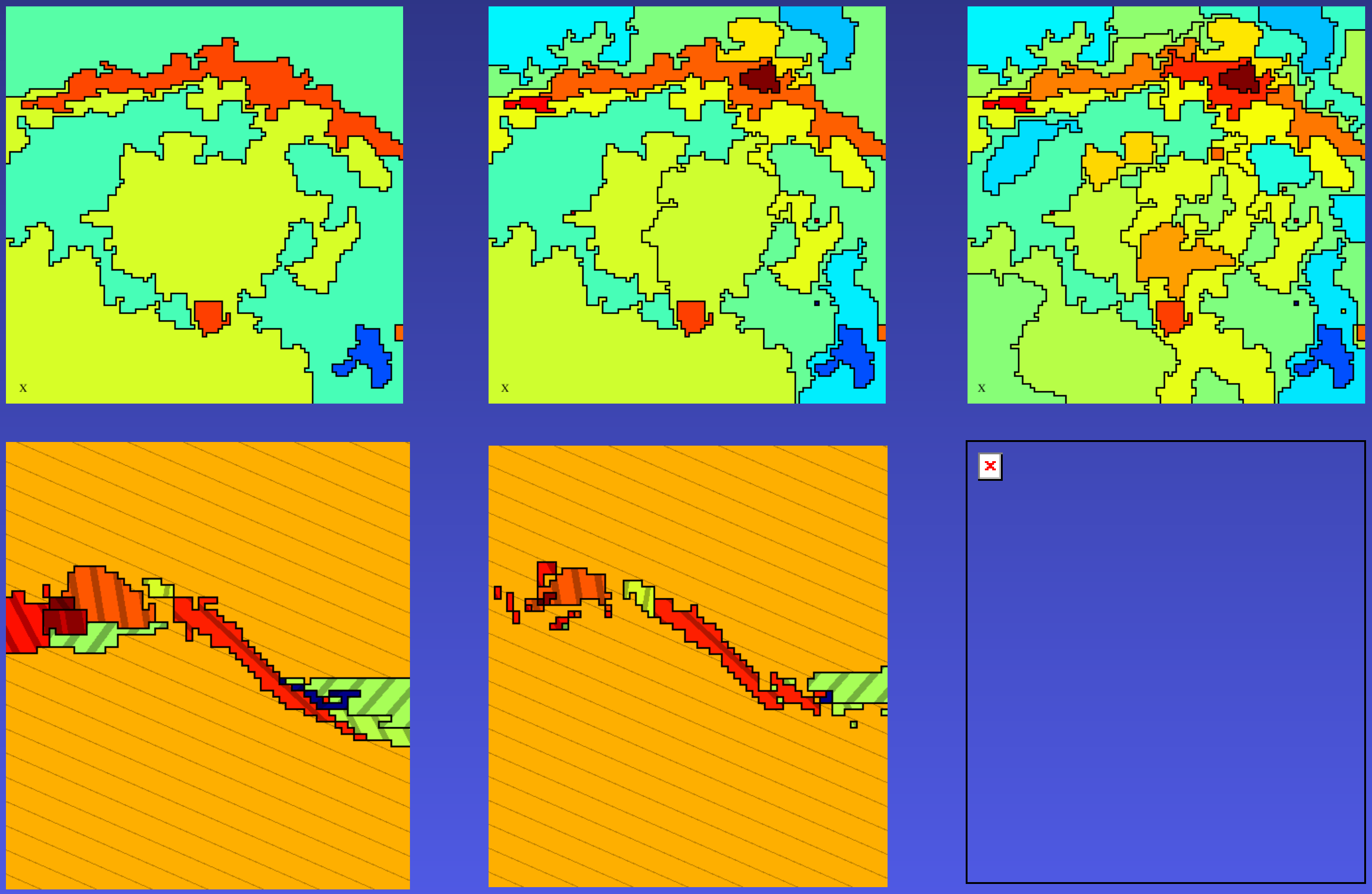


Streamline (white) trajectory of an ensemble forecast.



Pathline (yellow) trajectory of an ensemble forecast.

Clustering of Pixels



Top row: hierarchical clustering at different error thresholds (coarse to fine)

Bottom row: 3D clustering at varying depths. Texture patterns are used to denote positive and negative skewness of the distributions.

Publications

- "Visualizing Distributions from Multi-returned Lidar Data to Understand Forest Structures," D. Kao, M. Kramer, A. Luo, J. Dungan, and A. Pang, submitted to GeoInformatics 2004.
- “Visualization Techniques for Spatial Probability Function Data,” U. Borodoloi, D. Kao, H.-W. Shen, submitted to Data Science Journal, 2004.
- “Visualization and Exploration of Spatial Probability Density Functions: A Clustering Based Approach”, U. Borodoloi, D. Kao, H.-W. Shen, in Proc. of SPIE/ IS&T Conference on Visualization and Data Analysis, Jan. 2004.
- “Visualizing Spatial Distribution Data Sets,” A. Luo, D. Kao, and A. Pang, IEEE/Eurographics Visualization Symposium 2003, May 2003, pp.29-38.
- “Visualizing Spatially Varying Distribution Data”, D. Kao, A. Luo, J. Dungan, and A. Pang, Proc. of 6th International Conference on Information Visualization '02.
- “Visualizing Uncertainty in Geo-spatial Data,” A. Pang, Proceedings of the Workshop on the Intersections between Geospatial Information and Information Technology, National Academies committee of the Computer Science and Telecommunications Board, October 2001.
- “Visualizing 2D Probability Distributions from EOS Satellite Image-Derived Data Sets”, D. Kao, J. Dungan, and A. Pang, IEEE Visualization '01.

Extended Abstracts and Posters

- "Modeling and Visualizing Uncertainty in Continuous Variables Predicted Using Remotely Sensed Data," J. Dungan, D. Kao, and A. Pang, Proceedings of IGARSS '03.
- "Visualization of 2D Distributions from Models with Uncertainty," American Geophysical Union, EOS Trans. AGU, 83(47), Fall Meet. Suppl., B61A-0704, 2002.
- "Understanding Time-Varying Map Data Using Spatio-Temporal Clustering," U. Bordoloi, D. Kao, H.-W. Shen, EOS Trans. AGU, 83(47), Fall Meet. Suppl., NG12A-1018, 2002.
- "The Uncertainty Visualization Problem in Remote Sensing Analysis", J. Dungan, D. Kao, and A. Pang, Proceedings of IGARSS '02.

PDFVis Software

A software program for visualizing uncertainty, which can be represented by a probability density function (PDF) located at each grid cell in a spatial domain:

Main Features:

- The density estimate is interactively displayed at the pixel probe position.
- Parametric and non-parametric summaries of the distribution mean, median, quartiles, std. deviation, # of modes, and locations of modes are reported.
- Shaded surface displays of PDFs from a user-selectable profile (row or column) are projected onto a plane.

Linkable URL's

For more information about this research, see the following URL:

http://www.cse.ucsc.edu/research/avis/nasa_is